INCREASED PROTEIN EXPRESSION IN PLANTS

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application is a divisional of U.S. patent application Ser. No. 14/599,374, filed Jan. 16, 2015, now U.S. Pat. No. 10,774,336, issued Sep. 15, 2020, which claims the benefit of Patent Application Ser. No. 61/928,852 filed Jan. 17, 2014, the disclosure of which is hereby incorporated herein in its entirety by this reference.

FIELD OF THE DISCLOSURE

[0002] The present disclosure relates to methods and compositions for improving the expression of polynucleotides in a plant cell. In some embodiments, a protein encoding region of a polynucleotide is modified to reflect codon usage bias from a host organism while simultaneously preserving certain polyadenylation sequences in the native gene.

BACKGROUND

[0003] The genetic code consists of three-nucleotide units ("codons"). There are 64 possible codons, each specifying one of twenty amino acids, or an end to transcription (i.e., "STOP codons"). Therefore, at least some codons are redundant. In the coding system used by the vast majority of organisms, two amino acids are encoded by a single codon, whereas all other amino acids are encoded by two, three, four, or six codons, with three STOP codons. For amino acids with two, three, or four codons, the codons differ from each other at the third nucleotide position. For the three that have six codons, they have one block of four codons that follows this pattern, and one additional set of two that also differ from each other in the third position. For the two amino acids represented by six codons (Arg and Leu), they are each represented by a block different from the other by a change in the second nucleotide position. The codon representation of serine (Ser) is unusual, in that the two blocks of codons are very similar. For amino acids represented by two codons, the third position is either a purine (A, G) or pyrimidine (C, T) in both cases.

[0004] The degeneracy of the genetic code provides an opportunity to construct an alternative polynucleotide that encodes the polypeptide product of a reference polynucleotide. For example, codon degeneracy allows one to make synthetic DNA sequences that encode a protein of interest using codons that differ from those used in the original DNA coding sequence. For a particular amino acid, a given organism does not use the possible codons equally. Organisms each have a bias in codon usage. The pattern of bias in codon usage is distinct for an organism and its close relatives throughout the genome. For example, in *Streptomyces* spp., frequent codons generally include G or C in the third nucleotide position. Rare codons generally include A or T in the third position. In other organisms, A or T is preferred in the third position. Within a particular species, there can be distinct categories of genes with their own codon bias. In *E. coli*, for example, there are roughly three classes of genes, each with a distinctive codon usage signature. One class is rich in important proteins that are abundantly expressed; the second class includes proteins that are expressed at relatively low levels; and the third class includes proteins likely to have been recently acquired from other species.

[0005] To achieve desired expression levels of heterologous proteins in transgenic plants, it has been found beneficial to alter the native (sometimes referred to as wild-type or original) DNA coding sequence in various ways, for example, so that the codon usage more closely matches the codon usage of the host plant species, and/or so the G+C content of the coding sequence more closely matches the G+C level typically found in coding sequences of the host plant species, and/or so that certain sequences that destabilize mRNA are removed. For example, the expression of *Bacillus thuringiensis* (Bt) crystal protein insect toxins in plants has been improved using one or more of these approaches. See, e.g., U.S. Pat. Nos. 5,380,301; 5,625,136; 6,218,188; 6,340,593; 6,673,990; and 7,741,118.

[0006] In most synthetic gene design strategies, the process attempts to match the codon composition of a synthetic gene to the codon compositions of genes of a host in which the synthetic gene will be expressed. See, e.g., U.S. Patent Publication No. US 2007/0292918 A1. Such strategies may in some situations lead to increased expression of the synthetic gene in the host. For example, codon optimization in yeast may significantly improve the translation of heterologous gene transcripts due to minimizing the effects of, e.g., limiting aminoacyl-tRNAs and transcription termination at AT-rich sequences. See, e.g., Daly and Hearn (2004) J. Mol. Recognition 18:119-38.

[0007] However, despite general agreement in the art over the need for some sort of codon optimization, practitioners disagree over the general strategy that should be employed for optimization. One strategy that is preferred by some is to maximize the use of frequent codons in the expression host species during the design of heterologous genes. A second strategy preferred by others is to place maximum value on the context of particular codons, and therefore to maximize the use of codon pairs that occur frequently in the expression host. A third strategy is to make the codon usage of the new coding sequence in the new species resemble the codon usage of the reference coding sequence in the species of origin. This third strategy places high value on the recognition of possible requirements for rare codons to ensure proper secondary structure of transcript RNA molecules. Additionally, simply using the same frequently-occurring codon repeatedly in a heterologous sequence is expected to eventually have the same effect as selecting a rare codon; e.g., overuse of the corresponding tRNA will limit the availability of the tRNA. A person attempting to optimize the codons of a gene sequence for expression in a host organism must balance these strategies and their underlying concerns in order to arrive at a particular methodology.

[0008] The process of optimizing the nucleotide sequence coding for a heterologously expressed protein can be an important step for improving expression yields. However, several potential problems limit the usefulness of optimization for the expression of particular genes. For example, the secondary structure of an optimized transcript may limit translation of the transcript. Griswold et al. (2003) Protein Expression and Purification 27:134-42. Additionally, there are a number of sequence motifs that are desirably avoided in synthetic sequences for heterologous expression, including class I and II transcriptional termination sites in *E. coli* for a gene under the control of a T7 promoter; Shine-Dalgarno-like sequences; potential splice signals; sequences that promote ribosomal frameshifts and pauses; and polyadenylation signals. Welch et al. (2010) J. R. Soc. Interface